Parameter-Efficient Fine-Tuning of MedSAM for Prostate and Urethra Segmentation in Brachytherapy TRUS Images

Rucha Bhalchandra Joshi[1] R.JOSHI@CYI.AC.CY
Eleftherios Papageorgiou [1] E.PAPAGEORGIOU98@GMAIL.COM
Myrianthi Hadjicharalambous[1] M.HADJICHARALAMBOUS@CYI.AC.CY
Yiannis Roussakis[2] YIANNIS.ROUSSAKIS@GOC.COM.CY
Georgios Anagnostopoulos[2] GEORGIOS.ANAGNOSTOPOULOS@GOC.COM.CY
Iosif Strouthos[2] IOSIF.STROUTHOS@GOC.COM.CY
Constantinos Zamboglou[2] CONSTANTINOS.ZAMBOGLOU@GOC.COM.CY
Constantine Dovrolis[1,3] C.DOVROLIS@CYI.AC.CY
[1] *The Cyprus Institute, Nicosia, Cyprus*
[2] *German Oncology Center, European University Cyprus, Limassol, Cyprus.*
[3] *Georgia Institute of Technology, Atlanta, GA, USA*

Editors: Under Review for MIDL 2026

Abstract

Prostate and urethra segmentation in transrectal ultrasound (TRUS) images during brachytherapy is commonly performed manually, a process that is time-consuming, often exceeding 20 minutes - particularly when metallic brachytherapy needles introduce artifacts that obscure organ boundaries. The prolonged operating room time adds to staff burden and patient discomfort under anesthesia. Automated segmentation using medical foundation models such as MedSAM offers a direct solution, by reducing procedure time and improving brachytherapy workflow efficiency. Towards that aim, in this study we systematically evaluate 10 parameter-efficient fine-tuning strategies for MedSAM on 204 TRUS volumes, containing needle artifacts from brachytherapy procedures. The variants ranged from full retraining (100% parameters) to lightweight LoRA-based adaptations ($< 1\%$ parameters), targeting different architectural components of MedSAM.

The best-performing variant, MD Transformer LoRA, achieved a volume Dice score of 0.9484 [95% CI $0.9449, 0.9517$] for prostate segmentation and 0.9807 [95% CI $0.9800, 0.9813$] for urethra segmentation while training only 0.09% of model parameters. Parameter efficient variants consistently matched or exceeded full retraining performance across both in-house and three external datasets (3, 11, and 72 patients), substantially outperforming the original pretrained MedSAM (Dice: 0.8147) and nnU-Net baseline 0.8917. Based on our results, automated segmentation using parameter-efficient MedSAM fine-tuning can reliably replace manual delineation, reducing operation room time for the segmentation task from 20+ minutes to aruond 6 seconds per patient. This approach enables clinical deployment with minimal computational overhead while maintaining high accuracy (even in the presence of needle-induced artifacts), ultimately improving brachytherapy workflow efficiency and patient care. Our code is available at https://github.com/ruchajoshi/PROTECT.

Keywords: Parameter-efficient fine-tuning, MedSAM, Prostate brachytherapy, TRUS image segmentation, Foundation models for medical imaging

© 2026 CC-BY 4.0, R.B. Joshi *et al.*

1. Introduction

Prostate cancer is the second most frequent cancer and the fifth leading cause of cancer death among men, with over 1.4 million new cases annually worldwide (Sung et al., 2021). Prostate brachytherapy is a minimally invasive procedure for the treatment of prostate cancer, that involves the remotely driven insertion of a radioactive source (Iridium-192) through preimplanted needles to deliver dose inside the prostate gland. In this study, we primarily focus on the High-Dose-Rate Brachytherapy (HDR-BRT) that uses a higher dose rate over a shorter period.

A critical component of the workflow involves delineation of the prostate and urethra on transrectal ultrasound (TRUS) images to guide needle placement and dosimetry planning. Despite its necessity, this task remains largely manual, requiring experienced clinicians to segment anatomical structures slice-by-slice under time pressure. Segmentation alone can contribute significantly to overall operation room (OR) time, increasing procedure duration, staff burden, and patient discomfort. TRUS imaging during brachytherapy is particularly challenging because inserted needles create artifacts that occlude organ boundaries (figure 1), making delineation slow and potentially inaccurate. Automated segmentation would directly reduce OR time and improve brachytherapy workflow efficiency while also reducing inter-observer variability.

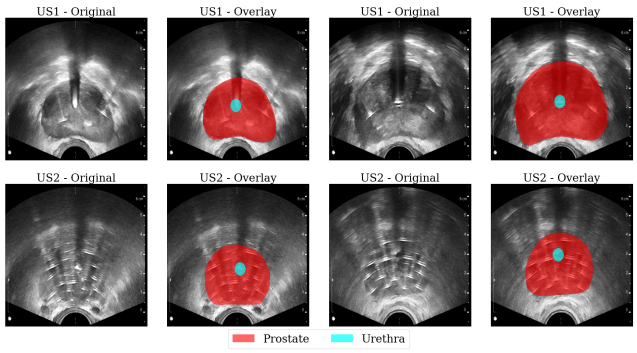

Figure 1: TRUS image examples with two anchor needles inserted in the prostate (top row) and with all needles inserted (bottom row) along with overlayed segmentation of prostate and urethra annotated by expert clinicians during the brachytherapy procedure.

Existing prostate segmentation methods for brachytherapy TRUS often require manual inputs such as seed points or landmarks (Zeng et al., 2018; Girum et al., 2020; Peng et al., 2023; Wang et al., 2025), requiring user expertise and limiting automation. Notably, most existing studies focus on artifact-free prostate images captured prior to needle insertion. Challenging segmentation tasks such as HDR-BRT prostate segmentation can be benefited by recent advances in large-scale medical foundation models, particularly those based on Segment Anything Model (SAM) (Kirillov et al., 2023), which have enabled reliable and efficient medical image segmentation. MedSAM (Ma et al., 2024), a medical adaptation of SAM pre-trained on large and diverse medical imaging datasets, providing strong general-

ization capabilities. It can be fine-tuned to adapt to new imaging conditions and unseen images, and thus offers an attractive option for HDR-BRT organ segmentation.

Despite the high potential of MedSAM, important open questions remain such as whether full model fine-tuning is necessary or if partial fine-tuning can yield similar performance. Parameter-efficient approaches are particularly advantageous in clinical environments where rapid adaptation to new imaging protocols is needed.

In this work, we address these gaps by introducing a reliable, accurate, and fast segmentation framework built upon the MedSAM foundation model. We enhance MedSAM's robustness through an optimized prompting strategy and systematically evaluate ten fine-tuning approaches, ranging from full retraining to parameter-efficient variants ($< 1\%$ trainable parameters). Our study leverages a unique, high-quality in-house dataset of 204 prostate cancer patients with expert-annotated ground-truth segmentations, a rare resource in this domain, and validates performance across three public datasets. The proposed method achieves state-of-the-art accuracy for both prostate and urethra segmentation, demonstrating that lightweight adaptations can match full retraining while enabling practical clinical deployment.

2. Related Work

Several recent methods have addressed the task of prostate segmentation in transrectal ultrasound (TRUS) images with varying levels of accuracy and reliance on user input. Girum et al. report a Dice similarity coefficient (DSC) of 96.9% for prostate segmentation; however, their method requires four user-provided input points, introducing a dependence on operator expertise. Peng et al. achieve a DSC of 96.2% by incrementally predicting the next vertex of the segmentation boundary given initial seed points. While accurate, this method also relies on manual interaction and was evaluated on TRUS images acquired without needle insertions, limiting its clinical realism. Wang et al. focus primarily on radiation plan quality prediction for high-dose-rate brachytherapy (HDRBT), incorporating prostate and organ segmentation as a subtask, and report a mean DSC of $90.0 \pm 8.0\%$. Hampole et al. utilize 3D TRUS data reformatted into multiple 2D planes to train U-Net and V-Net models, reporting a median DSC of 87.2% [interquartile range: 84.1–88.8%].

Most existing methods require manual seed points or initial landmarks, introducing user dependency and limiting automation. Furthermore, few works address segmentation in the presence of needle artifacts common during intra-operative brachytherapy imaging. Our approach eliminates manual interaction by using automated bounding box prompts and is specifically adapted to handle needle-induced artifacts.

Urethra segmentation in TRUS imaging for brachytherapy remains a rarely addressed task in the literature, despite its clinical importance for treatment planning and dose optimization. The small anatomical size of the urethra and low contrast relative to surrounding prostate tissue make automated segmentation particularly challenging. To our knowledge, no prior work has systematically evaluated foundation model adaptation for automated urethra segmentation in TRUS brachytherapy workflows. The scarcity of urethra-specific segmentation methods highlights a critical gap in the field.

Recent developments in large-scale medical foundation models, particularly those based on the Segment Anything Model (Kirillov et al., 2023), have opened avenues for reliable and

efficient medical image segmentation. MedSAM (Ma et al., 2024), a medical adaptation of SAM, is pre-trained on over 1.5 million medical image-mask pairs across diverse modalities and anatomical structures, providing strong generalization capabilities. Unlike earlier task-specific segmentation models, foundation models can be fine-tuned with minimal data and adapt to new imaging conditions with substantially fewer annotated examples. Parameter-efficient fine-tuning strategies, such as Low-Rank Adaptation (LoRA) (Hu et al., 2022) and adapter modules, have emerged as effective approaches to adapt large pre-trained models while training only a small fraction of parameters, reducing computational costs and mitigating overfitting risks.

3. Methods

3.1. Datasets

We acquired a high-quality in-house dataset from the patients undergoing the brachytherapy procedure at the German Oncology Center, Limassol, Cyprus. This dataset contains the TRUS images collected at different times during the brachytherapy procedure.

Table 1: Summary of the datasets used in this study for fine-tuning and evaluation of MedSAM variants. We use TRUS images for all datasets.

Dataset	Number of Patients	Needle Inserted	Usage	Annotated by
In-house GOC dataset	204	Yes	fine-tuning + Testing	Expert researchers and clinicians
mu-Reg Challenge Dataset (Baum et al., 2023)	72	No mention	Testing	Expert researchers and clinicians
Annotated MRI and Ultrasound volume images of the prostate (Fedorov et al., 2015)	3	No mention	Testing	Clinicians
Prostate MRI and Ultrasound from Kaggle (Natarajan et al., 2020)	11/1151 [1]	No mention	Testing	Semi-automatic

The TRUS images are acquired at two key stages of the brachytherapy procedure. First, images are recorded immediately after two anchor needles are implanted in the prostate to stabilize it during planning and needle insertion phase. These images are designated as 'US1'. Following this initial imaging, needles are strategically placed throughout the prostate to maximize coverage while minimizing radiation exposure to organs at risk (OARs). The second set of TRUS images, designated as 'US2', is acquired after all needles have been placed. Example slices are shown in figure 1. For a particular patient, organ in US2 images might be slightly deformed compared to US1 images due to the insertion of multiple needles. These images contain significant artifacts caused by needles, which obscure the boundaries of both the prostate, urethra and surrounding OARs. Expert clinicians manually annotate the segmentations in both US1 and US2 images, providing ground truth labels for model fine-tuning.

This in-house GOC dataset consists of 204 patients' data in total. We make a 70%/15%/15% train/validation/test splits using different seeds to fine-tune and evaluate our model variants. Apart from this in-house dataset, we validate our results on three other datasets that

1. We sample 11 patients' TRUS volumes out of 1151.

contain TRUS images of prostate, along with the annotated ground truth. The details of all the datasets used in this study are summarized in table 1.

3.2. Models

3.3. Baseline Models

We used the pretrained MedSAM model as our primary foundation model baseline. The base MedSAM model consists of an Image Encoder (IE), Prompt Encoder (PE), and Mask Decoder (MD). It is pre-trained on a large-scale medical image dataset consisting of more than 1.5 million medical image-mask pairs. This dataset comprised of 10 imaging modalities and over 30 cancer types and is designed to enable quick adaptation to downstream segmentation tasks with limited labeled data.

As another strong segmentation baseline, we trained nnUNet (Isensee et al., 2021) on the same training set using its standardized pipeline. nnUNet serves as a widely recognized benchmark due to its proven ability to adapt to diverse medical imaging modalities and tasks.

3.4. MedSAM fine-tuning variants

To investigate the impact of parameter-efficient adaptation, we designed 10 fine-tuning variants of MedSAM, each defined by a different combination of frozen and trainable blocks or layers. We evaluated them on in-house transrectal ultrasound (TRUS) prostate dataset. Each strategy differs in which modules were retrained or frozen, and whether low-rank adapters (LoRA) or bottleneck adapters were inserted into the Mask Decoder.

All methods were compared on identical data splits and evaluation metrics to isolate the impact of the adaptation strategy. The model variants are described in the table 2 that gives details regarding the parts of the MedSAM architecture that are trainable in each variant.

Table 2: Description of MedSAM variants, showing the total and trainable parameters, the percentage of trainable parameters along with a brief description of the trainable parts of the MedSAM model.

Model Variant	Trainable blocks in model architecture	Total Parameters	Trainable Parameters	Percentage
Full retraining	All	93735472	93735472	100%
Partial IE	Last 2 blocks of IE, MD	93735472	18253796	19.47%
Freeze IE	PE, MD	93735472	4064560	4.34%
MD only	MD	93735472	4058340	4.33%
Selective layers	Last 2 transformer blocks and iou, hyperparam heads in MD	93735472	3850884	4.11%
Hybrid LoRA	LoRA on attention layers in MD transformer blocks + fine-tuning of classifier heads.	93817392	773764	0.82%
MD Hypernetwork MLP	Only classifier heads in MD are fine-tuned.	93735472	691844	0.74%
MD Lora	LoRA (r=8) to transformer(attention + MLP)	93891120	155648	0.17%
MD Transformer LoRA	LoRA (r=8) on Q/K/V (attention layers) and not MLPs in MD transformers	93817392	81920	0.09%
MD MLP LoRA	LoRA (r=8) applied solely to MLP sublayers of MD transformer blocks.	93809200	73728	0.08%

3.5. Prompting Strategy

MedSAM uses prompts to control segmentation outputs. In this study, we focus on bounding box prompts to delineate the prostate and urethra. All models received identical prompt inputs so that performance differences reflect model adaptation rather than prompt variation.

The default perturbation used by the original MedSAM implementation is as follows: Let $r_i \sim \mathcal{U}(0, s)$ for $i = 1, 2, 3, 4$ where s is the maximum shift. This value of maximum shift is specified in pixels. Then the augmented bounding box coordinates are:

$$\begin{aligned} x'_{\min} &= \max(0, x_{\min} - r_1) & x'_{\max} &= \min(W, x_{\max} + r_2) \\ y'_{\min} &= \max(0, y_{\min} - r_3) & y'_{\max} &= \min(H, y_{\max} + r_4) \end{aligned} \tag{1}$$

where $x'_{min}, x'_{max}, y'_{min}, y'_{max}$ are minimum and maximum values of x and y coordinates as indicated, W and H are the width and height of the entire image slice and $\mathcal{U}$ indicates the uniform distribution with given parameters. However, this perturbation (equation 1) to the bounding boxes does not cover all the possible bounding boxes around the organ to be segmented. The real world prompts given by the clinicians may vary in position, size, and aspect ratio.

For prostate segmentation, to make our model robust to any possible bounding box around the organ, a different perturbation was used. This is described as follows: The lower-left corner of the bounding box, (x_{min}, y_{min}) is treated as a fixed anchor point. The coordinates and the length of the diagonal are modified using multiplicative perturbations, as defined below:

$$\begin{aligned} x'_{min} &= x_{min} \cdot (1 + \epsilon_x) \\ y'_{min} &= y_{min} \cdot (1 + \epsilon_y) \\ l' &= l \cdot (1 + \epsilon_d) \end{aligned} \tag{2}$$

where $\epsilon_x, \epsilon_y, \epsilon_l$ are drawn from a uniform distribution $\mathcal{U}(-\epsilon, \epsilon)$, ϵ denotes maximum allowed perturbation, and l represents the diagonal length of the bounding box. The upper-right corner (x'_{max}, y'_{max}) is then calculated based on the adjusted anchor point and the modified diagonal, depending on the desired orientation and aspect ratio. We also experimented with $\mathcal{U}(0, \epsilon)$ to analyse the comparative model performance.

For urethra segmentation, experiments were performed with only enlargement of the bounding boxes in case of urethra segmentation. For enlargement the centre of the bounding box acts as an immovable anchor, and a $\mathcal{U}(0, \epsilon)$ scaling is applied to the length of diagonal. We observed that this strategy works better for urethra segmentation than the one in equation 2 as the organ is very small and any shift in the bounding box may lead to missing the organ completely.

3.6. Model Optimization

We fine-tuned all models using the AdamW optimizer with weight decay of $1e-2$. The base learning rate was set to $1e-4$ for fully trainable components similar to the original MedSAM; for adapter- and LoRA-only runs, we applied a higher effective learning rate of $1e-3$ to the small adapter parameters. We minimized an equally weighted sum of Dice loss

and cross-entropy. Training used a per-GPU batch size of 2 volumes. Models were trained for up to 200 epochs with early stopping triggered when validation Dice failed to improve for 20 consecutive epochs.

4. Evaluation Metrics

We evaluate all the model variations on several metrics. Primarily, to get an overall evaluation of predicted segmentation for every patient, we consider the slice-level and volume-level metrics. The slice-level metrics include Dice coefficient, Jaccard index, Dice median, Precision and Recall. The slice level metrics allow us to analyze the model behavior across the apex, mid-gland, and base of the prostate. The volume-level metrics reported are Dice coefficient, Jaccard index, 95th percentile Hausdorff distance (HD95) in mm, Precision and Recall. Volume-level metrics give clinically relevant performance measure for structure-level delineation used in treatment planning.

To measure the efficiency of the model, we consider the trainable parameter count (absolute and percentage of full model). This metric helps in analysis of parameter-efficient fine-tuning strategies that may be beneficial for quick clinical deployment in different settings.

5. Results

5.1. Performance on the in-house dataset (GOC Dataset) - Volume Level

The volume-level segmentation performance of different MedSAM fine-tuning strategies on the in-house GOC dataset is summarized in table 3. The MedSAM variant MD Transformer LoRA achieves the highest volume-level Dice score of 0.948 ± 0.003 . This variant applies LoRA adapters exclusively to the attention layers of the Mask Decoder transformers, leaving MLP layers frozen. This suggests that adapting attention mechanisms is particularly effective for capturing relevant features in TRUS images for prostate segmentation. Full Retraining and Partial IE achieve lower HD95 scores of 5.663 and 5.727 respectively, MD Transformer LoRA achieves a competitive HD95 of 5.866 ± 0.737 while training only 0.09% of the parameters. The strong performances of models such as Full Retraining and Partial IE requires updating high number of model parameters. Other low-parameter variants also demonstrate competitive performance as shown in table 3. The relatively small performance gap between the best (MD Transformer LoRA) and worst (MD Hypernetwork MLP) fine-tuning strategies suggests that MedSAM's pretrained features are broadly useful for TRUS segmentation when appropriately fine-tuned. All fine-tuned MedSAM variants show significant improvement across all metrics compared to the original pretrained MedSAM. Notably, all fine-tuned variants outperform both the original MedSAM and the nnU-Net baseline.

5.2. Performance on the in-house dataset (GOC Dataset) - Slice Level

MD Transformer LoRA achieved the best slice-level performance with a Dice score of 0.941 ± 0.004, Jaccard score of 0.890 ± 0.006while training only 0.09% of the parameters. Parameter-efficient variants including Hybrid LoRA (0.082% trainable, Dice 0.939) and MD LoRA (0.17% trainable, Dice 0.937) demonstrated competitive performance to full retraining. All

Table 3: Volume-level metrics with 95% bootstrap confidence intervals for different MedSAM fine-tuning strategies on GOC prostate segmentation. The best scores are highlighted in bold, and the second best are underlined.

Model Variant	Trainable %	Dice	Jaccard	Precision	Recall	HD95
Full Retraining	100.00%	$\underline{0.947 \pm 0.004}$	$\underline{0.900 \pm 0.008}$	$\underline{0.955 \pm 0.009}$	$\underline{0.941 \pm 0.011}$	$\mathbf{5.663 \pm 0.899}$
Partial IE	19.47%	0.944 ± 0.005	0.894 ± 0.009	0.954 ± 0.009	0.935 ± 0.014	$\underline{5.727 \pm 0.908}$
Freeze IE	4.34%	0.938 ± 0.008	0.883 ± 0.013	0.952 ± 0.010	0.925 ± 0.017	6.950 ± 1.293
MD Only	4.33%	0.940 ± 0.006	0.887 ± 0.010	0.939 ± 0.013	$\underline{0.942 \pm 0.009}$	7.628 ± 1.206
Selective Layers	4.11%	0.941 ± 0.005	0.888 ± 0.009	0.942 ± 0.012	0.940 ± 0.010	6.719 ± 0.835
Hybrid LoRA	0.82%	0.944 ± 0.005	0.899 ± 0.008	$\mathbf{0.955 \pm 0.008}$	0.940 ± 0.010	6.022 ± 1.238
MD Hypernetwork MLP	0.74%	0.914 ± 0.006	0.842 ± 0.010	0.911 ± 0.009	0.917 ± 0.009	17.901 ± 2.712
MD LoRA	0.17%	0.946 ± 0.004	0.897 ± 0.008	0.954 ± 0.008	0.938 ± 0.013	6.439 ± 0.913
MD Transformer LoRA	0.09%	$\mathbf{0.948 \pm 0.003}$	$\mathbf{0.902 \pm 0.007}$	0.952 ± 0.006	$\mathbf{0.945 \pm 0.007}$	5.866 ± 0.737
MD MLP LoRA	0.08%	0.938 ± 0.004	0.884 ± 0.007	0.954 ± 0.008	0.923 ± 0.013	7.162 ± 0.881
MedSAM Original	0.00%	0.815 ± 0.016	0.689 ± 0.022	0.929 ± 0.007	0.727 ± 0.024	16.256 ± 1.847
nnU-Net	-	0.898 ± 0.013	0.806 ± 0.021	0.873 ± 0.019	0.914 ± 0.012	7.170 ± 0.790

fine-tuned MedSAM variants substantially outperformed the original pretrained MedSAM (Dice 0.801) and nnU-Net (Dice 0.836), demonstrating that domain adaptation is essential for TRUS segmentation with needle-induced artifacts. Detailed results are provided in table 8 in A.

5.3. Region-wise and Slice-position analysis

Segmentation accuracy varied systematically across prostate regions for all models. The mid-gland consistently achieved the highest Dice scores across all variants. We attribute this to the higher density of positive pixels in the ground truth segmentation at the mid-gland compared to the base and apex regions. The slice-position versus Dice curve demonstrated a consistent pattern where performance peaks at the mid-gland and declines toward the apex and base. Figure 2 illustrates these trends for MD Transformer LoRA, which remain consistent across all fine-tuned variants.

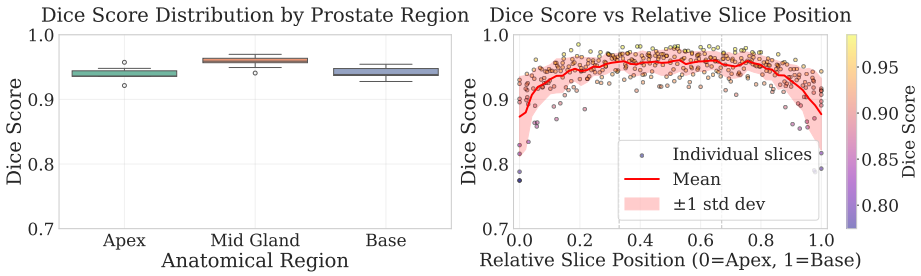

Figure 2: Slice-wise performance of the model MD Transformer LoRA.

5.4. Qualitative Evaluation

Fine-tuned MedSAM variants produced smooth boundaries substantially outperforming the original pretrained MedSAM. Figure 3 compares pretrained MedSAM with MD Transformer

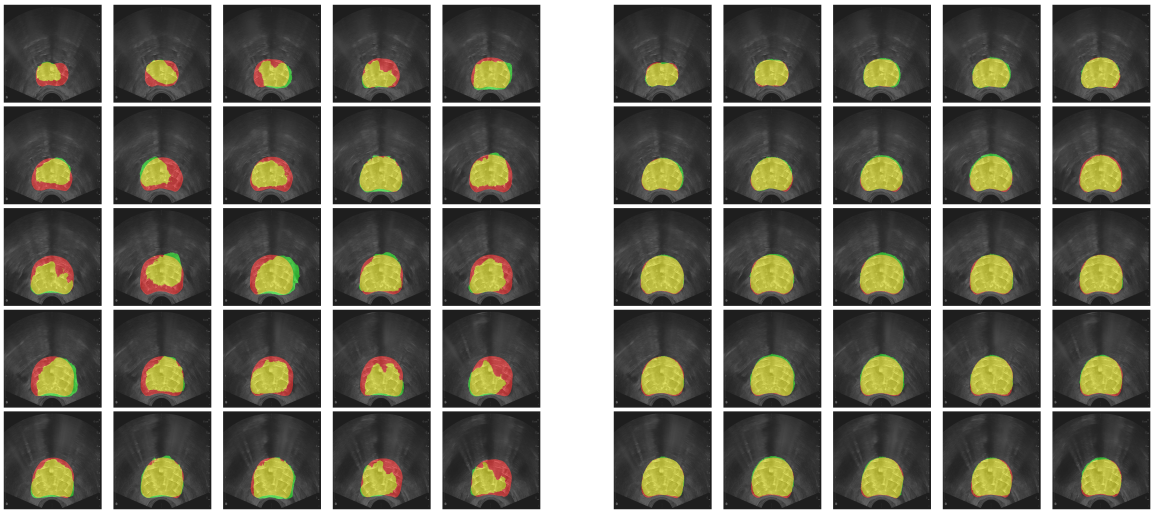

Figure 3: Output of the pretrained MedSAM (left) and MD Transformer LoRA (right) for different slices of a single patient from the test set with needles inserted in the prostate. The ground truth annotation is shown in red (False Negative), model prediction is shown in green (False Positive), and the overlap between the ground truth and the model prediction is shown in yellow (True Positive).

LoRA on the same test TRUS volume, showing pretrained model struggles with needle artifacts leading to significant false negatives and positives, while the fine-tuned variant accurately delineates boundaries. Figure 4 shows the worst performing slices where the model still captures overall prostate shape despite significant needle artifacts, with failures occuring mainly at apex/base regions or shadow-heavy slices.

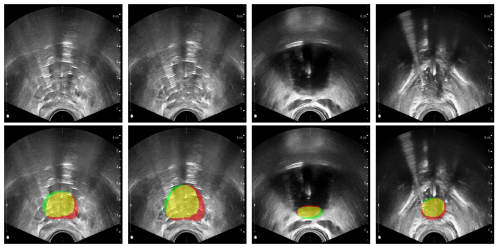

Figure 4: Output of the best performing MedSAM variant MD Transformer LoRA for the worst 4 segmentation slices from the test set with needles inserted in the prostate. The top row shows the TRUS images with needles, the bottom row shows the corresponding ground truth and predicted segmentations. The ground truth annotation is shown in red (False Negative), model prediction is shown in green (False Positive), and the overlap between the ground truth and the model prediction is shown in yellow (True Positive).

5.5. Performance Comparison by TRUS Acquisition Stage

Table 4: Comparison of segmentation performance between TRUS acquisition stages with different levels of needle artifacts.

	US1 (TRUS with anchor needles)	US2 (TRUS with all needles)
Dice	0.953 ± 0.007	0.946 ± 0.008
HD95	4.577 ± 1.443	6.262 ± 1.417

To assess model robustness across different levels of needle artifacts, we compared the performance of the best-performing variant (MD Transformer LoRA) on US1 (TRUS with anchor needles only) versus US2 (TRUS with all needles inserted) acquisitions. As shown in table 4, the model achieved a Dice score of 0.953 ± 0.007 on US1 images and 0.946 ± 0.008 on US2 images, demonstrating minimal performance degradation despite increased artifact complexity. The HD95 increased from 4.577 ± 1.443 mm on US1 to 6.262 ± 1.417 mm on US2, indicating slightly reduced boundary precision with additional needle artifacts. These results demonstrate that the fine-tuned model maintains robust performance even under challenging imaging conditions with extensive needle-induced artifacts, supporting its clinical viability for intra-operative segmentation throughout the brachytherapy workflow.

5.6. Urethra Segmentation Performance

Urethra segmentation poses a distinct challenge in TRUS-guided brachytherapy due to its small size within the prostate volume and lower contrast relative to surrounding prostate tissue. Despite these challenges, fine-tuned MedSAM variants substantially improve upon the baseline MedSAM, which performs poorly at both slice and volume levels. For slice-level segmentation, the original MedSAM achieved a Dice score of 0.6792 ± 0.0078 compared to the best fine-tuned variant at 0.9804 ± 0.0008 . At the volume level, the improvement was similarly dramatic: 0.6857 ± 0.0084 for original MedSAM versus 0.9807 ± 0.0006 for the best fine-tuned variant. These results demonstrate the effectiveness of fine-tuning strategies for urethra segmentation in TRUS images. Consistent improvements across all metrics are shown in tables 5 and 9 (appendix) , confirming that task-specific adaptation of MedSAM is essential for reliable urethra segmentation.

A key observation is that most of the fine-tuning methods give dice scores around 0.978, which suggests that the urethra is highly learnable even with minimal fine-tuning. Secondly, the parameter efficient model variations either match or outperform the full retraining of MedSAM.

5.7. Evaluation on Public Datasets

A summary of the volume-level Dice and HD95 results on the three public datasets after fine-tuning on the GOC dataset is presented in table 6. MD MLP LoRA emerged as the best-performing parameter-efficient variant across all three public datasets, achieving the highest Dice scores of 0.862 ± 0.006, 0.874 ± 0.020, and 0.859 ± 0.022 on the 3-patient, 11-patient, and 72-patient datasets respectively, with MD Hypernetwork MLP and Hybrid LoRA also demonstrating strong generalization. These parameter-efficienct variants ($< 1\%$

Table 5: Volume-level performance metrics (95% bootstrap CI) for different model variants on urethra segmentation on in-house GOC dataset. The best scores are highlighted in bold, and the second best are underlined.

Model Variant	Trainable %	Dice	Jaccard	Precision	Recall	HD95
Full Retraining	100.00	0.9783 ± 0.0016	0.9576 ± 0.0027	0.9797 ± 0.0029	0.9771 ± 0.0017	1.0573 ± 0.0628
Partial IE	19.47	0.9797 ± 0.0013	0.9603 ± 0.0024	0.9796 ± 0.0021	0.9799 ± 0.0010	1.0363 ± 0.0545
Freeze IE	4.34	0.9786 ± 0.0015	0.9581 ± 0.0025	0.9754 ± 0.0025	$\mathbf{0.9819 \pm 0.0016}$	1.0512 ± 0.0526
MD Only	4.33	0.9794 ± 0.0018	0.9597 ± 0.0032	0.9800 ± 0.0029	0.9789 ± 0.0014	1.0929 ± 0.1081
Selective Layers	4.11	0.9789 ± 0.0019	0.9588 ± 0.0037	0.9774 ± 0.0039	$\underline{0.9806 \pm 0.0014}$	1.0743 ± 0.0960
Hybrid LoRA	0.82	0.9800 ± 0.0007	0.9609 ± 0.0013	0.9796 ± 0.0021	0.9806 ± 0.0016	$\mathbf{1.0078 \pm 0.0117}$
MD Hypernetwork MLP	0.74	0.9028 ± 0.0024	0.8231 ± 0.0040	0.8801 ± 0.0044	0.9274 ± 0.0044	4.5419 ± 0.1894
MD LoRA	0.17	$\underline{0.9804 \pm 0.0009}$	$\underline{0.9615 \pm 0.0016}$	$\underline{0.9822 \pm 0.0017}$	0.9786 ± 0.0013	1.0184 ± 0.0238
MD Transformer LoRA	0.09	$\mathbf{0.9807 \pm 0.0007}$	$\mathbf{0.9621 \pm 0.0012}$	$\mathbf{0.9831 \pm 0.0015}$	0.9784 ± 0.0015	$\mathbf{1.0078 \pm 0.0117}$
MD MLP LoRA	0.08	0.9780 ± 0.0006	0.9570 ± 0.0011	0.9765 ± 0.0013	0.9796 ± 0.0013	1.0088 ± 0.0110
Original MedSAM	0.00	0.6857 ± 0.0085	0.5232 ± 0.0098	0.9733 ± 0.0022	0.5313 ± 0.0100	10.2937 ± 0.3612

Table 6: Volume-level Dice and HD95 results on public datasets after fine-tuning on GOC dataset.

Dataset	Model	Volume Level Dice	Volume Level HD95
3 patient	Full retraining	0.825 ± 0.024	71.41 ± 63.64
	Best parameter-efficient variant (MD MLP LoRA)	$\mathbf{0.862 \pm 0.006}$	9.09 ± 2.30
	Pretrained MedSAM	0.836 ± 0.035	$\mathbf{4.08 \pm 1.54}$
11 patient	Full retraining	0.851 ± 0.027	64.67 ± 12.62
	Best parameter-efficient variant (MD MLP LoRA)	$\mathbf{0.874 \pm 0.020}$	50.16 ± 6.84
	Pretrained MedSAM	0.851 ± 0.012	$\mathbf{38.41 \pm 3.35}$
72 patient	Full retraining	0.782 ± 0.050	92.91 ± 33.85
	Best parameter-efficient variant (MD MLP LoRA)	$\mathbf{0.859 \pm 0.022}$	43.09 ± 15.70
	Pretrained MedSAM	0.793 ± 0.038	$\mathbf{26.88 \pm 7.13}$

parameters) consistently outperformed full retraining, sugesting selective adaptation prevents overfitting while preserving generalizable representations. A striking HD95 paradox emerged: pretrained MedSAM achieved the lowest HD95 values (4.08 ± 1.54, 38.41 ± 3.35, 26.88 ± 7.13) across datasets despite lower Dice scores, indicating fewer extreme boundary errors even with overall poorer segmentation quality. Conversely, full retraining exhibited highly variable HD95 (71.41 ± 63.64, 64.67 ± 12.62, 92.91 ± 33.85), reflecting unstable boundary predictions. For brachytherapy workflows requiring complete organ delineation, parameter-efficient variants remain optimal despite occasional boundary outliers. The detailed results on the external datasets are given in appendix C.

5.8. Parameter Efficiency

A key finding is that the high accuracy segmentation does not require full fine-tuning. Variants training only $0.8\% - 4\%$ of parameters achieved performance comparable and indistinguishable from the full model. We present the visual comparison of the different fine-tuning strategies with the baseline in figure 5 that also highlights the number of parameters trained for the particular model.

Through fine-tuning various MedSAM model variants, we assess the significance of fine-tuning different blocks or layers by analysing their performance on evaluation metrics de-

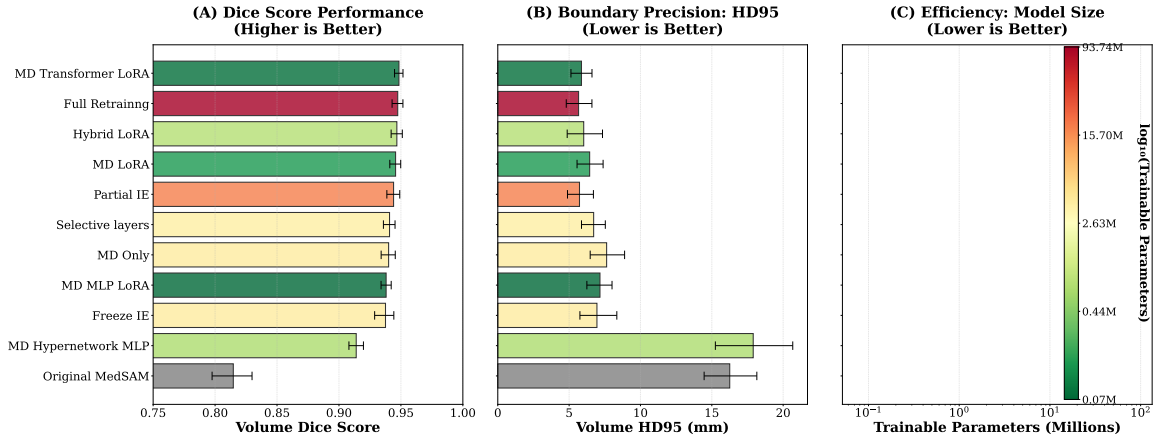

Figure 5: Performance, boundary precision, and efficiency comparison of MedSAM variants. Higher Dice and lower HD95 scores are achieved with fewer trainable parameters.

scribed in Section 4. When all MedSAM parameters are fine-tuned on the GOC specific dataset, performance improves relative to the original MedSAM model. However, other MedSAM variants demonstrate superior or comparable performance across various evaluation metrics while fine-tuning substantially fewer parameters, as illustrated in figure 5. Given the relatively small dataset size compared to the total parameter count, fine-tuning all parameters may not be optimal and could lead to overfitting.

For the model MD Transformer LoRA, we add LoRA adapters to the query, key, and value projection layers of the transformer blocks in the Mask Decoder (MD), without altering the MLP parameters in MD. This allows the model to adapt the attention mechanism to better capture the relationships between image features and prompts specific to TRUS images with needle artifacts. The performance of this model is the best among all the variants, though a significantly small number of parameters are fine-tuned. The model efficiently learns the semantics of the training data that comprises of images with and without artifacts added by needles in the TRUS images, unlike in case of the training data for the baseline MedSAM. The inference times for some of the MedSAM variants are given in table 7. Though MedSAM has a very small inference time, the parameter-efficient fine-tuning methods further reduces the inference time slightly making it more suitable for clinical deployment.

Table 7: Inference times for different MedSAM variants on the GOC dataset.

Model	per Volume Time (s)	per Slice Time (s)	Frames per second
Full retraining	6.09 ± 1.32	0.261 ± 0.059	3.83
MD Transformer LoRA	5.62 ± 0.84	0.248 ± 0.040	4.04
Pretrained MedSAM	6.12 ± 1.24	0.263 ± 0.065	3.80

6. Discussion

The pairwise Wilcoxon signed-rank test results (figure 6) reveal that parameter-efficient fine-tuning variants, particularly MD Transformer LoRA, achieve statistically equivalent performance to full retraining while using only 0.09% of model parameters. This finding provides strong evidence that MedSAM's attention mechanisms in the mask decoder are the most critical components for domain adaptation to TRUS imaging with needle artifacts. The superior performance of transformer-based adaptations (MD Transformer LoRA, Hybrid LoRA) compared to MLP-only approaches suggests that attention layers more effectively capture domain-specific feature relationships essential for distinguishing organ boundaries amid metallic artifacts.

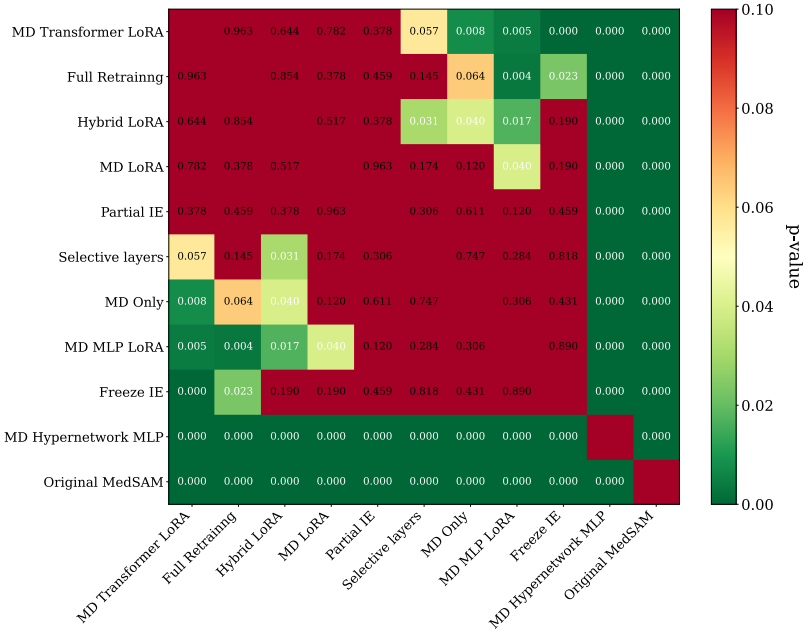

Figure 6: Pairwise Wilcoxon test p-values for MedSAM variants. Green cells indicate significant difference ($p < 0.05$); red cells indicate no significant difference.

Notably, all fine-tuned variants significantly outperformed the original MedSAM ($p < 0.05$), reinforcing that domain-specific adaptation is essential for clinical deployment. The failure of MD Hypernetwork MLP relative to other variants indicates that adapting only classifier heads without modifying feature extraction mechanisms is insufficient for this challenging segmentation task. These technical insights demonstrate that foundation models can be efficiently adapted for specialized medical tasks without the computational overhead and overfitting risks associated with full retraining, enabling practical deployment in resource-constrained clinical environments. Our statistical and multi-seed analyses confirm the robustness required for clinical deployment, moving this method beyond proof-of-concept to validation-level evidence.

6.1. Clinical Impact

From a clinical perspective, the results indicate that the prostate and urethra segmentation during brachytherapy can be automated reliably using MedSAM based models. This is applicable even in the cases where the needles are implanted during the TRUS imaging. This automation would significantly reduce the operation room time, the duration for which the patient is under the effect of anesthesia. In a clinical setting, prostate and urethra segmentation even when performed by an expert clinician can take around 20 minutes or more per patient. The automated segmentation can reduce this time to a few seconds ($\sim$ 6 seconds), thus improving the overall workflow of the brachytherapy procedure. The automation also avoids the variations introduced by different expertise level of the clinical annotator.

6.2. Limitations

While this study demonstrates the significance of fine-tuning methodologies across different datasets, there are several limitations to it. First, our bounding box prompts were derived from ground-truth annotations, representing idealized conditions, clinical deployment would require automated localization methods. Second, urethra segmentation evaluation was limited to our in-house dataset due to scarcity of publicly available TRUS datasets with urethra annotations, limiting comprehensive validation. Third, the external datasets did not explicitly document needle artifact presence during acquisition, making it unclear whether generalization was tested specifically under similar artifact conditions as our in-house data.

Future work will adress these limitations by developing automated organ localization methods for bounding box generation and evaluating their impact on segmentation performance. Multi-institutional prospective studies with standardized urethra annotations would enable more comprehensive validation. Additionally, systematic investigation of multi-organ segmentation (including bladder and rectum) would further establish clinical viability. Finally, integration into real-time clinical workflows with user feedback would help refine the system for practical deployment in brachytherapy procedures.

7. Conclusion

This study demonstrates that foundation models can be effectively adapted for challenging clinical tasks through parameter efficient fine-tuning, with our best variant (MD Transformer LoRA) achieving Dice scores of 0.948 (prostate) and 0.9787 (urethra) while training only 0.09% of the parameters. This parameter efficiency enables deployment with minimal computational resources, addressing a key barrier to adopting foundation models in resource-constrained healthcare environments.

Automated segmentation significantly reduces OR time during TRUS-guided prostate brachytherapy, improving clinical workflow and patient experience. Consistent performance across multiple external datasets despite varying imaging conditions shows the generalization essential for clinical adoption. Our systematic evaluation of 10 fine-tuning strategies provides actionable guidance for adapting foundation models to medical imaging tasks with limited data, contributing evidence that parameter efficient fine-tuning can accelerate the translation of AI advances for better clinical impact.

Acknowledgments

This work was supported through the project CODEVELOP-AG-SH-HE/0823/0114 - funded by the European Union NextGenerationEU instrument, through the Research and Innovation Foundation of Cyprus.

References

Zachary M. C. Baum, Shaheer U. Saeed, Zhe Min, Yipeng Hu, and Dean C. Barratt. Mr to ultrasound registration for prostate challenge - dataset (1.1.0), 2023. [Data set].

A. Fedorov, S. Khallaghi, C. Antonio Sánchez, A. Lasso, S. Fels, K. Tuncali, Sugar, E. N., T. Kapur, C. Zhang, W. Wells, Nguyen, P. L., P. Abolmaesumi, and C. Tempany. Open-source image registration for MRI–TRUS fusion-guided prostate interventions. *International Journal of Computer Assisted Radiology and Surgery*, 10(6):925–934, 2015. doi: 10.1007/s11548-015-1180-7.

Kibrom Berihu Girum, Gilles Créhange, Raabid Hussain, and Alain Lalande. Fast interactive medical image segmentation with weakly supervised deep learning method. *International Journal of Computer Assisted Radiology and Surgery*, 15(9):1437–1444, 2020.

Prakash Hampole, Thomas Harding, Derek Gillies, Nathan Orlando, Chandima Edirisinghe, Lucas C Mendez, David D'Souza, Vikram Velker, Rohann Correa, Joelle Helou, et al. Deep learning-based ultrasound auto-segmentation of the prostate with brachytherapy implanted needles. *Medical Physics*, 51(4):2665–2677, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024.

S. Natarajan, A. Priester, D. Margolis, J. Huang, and L. Marks. Prostate MRI and Ultrasound with pathology and coordinates of tracked biopsy (prostate-mri-us-biopsy), 2020. [Data set].

Tao Peng, Yan Dong, Gongye Di, Jing Zhao, Tian Li, Ge Ren, Lei Zhang, and Jing Cai. Boundary delineation in transrectal ultrasound images for region of interest of prostate. *Physics in Medicine & Biology*, 68(19):195008, 2023.

Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021. doi: https://doi.org/10.3322/caac.21660. URL https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660.

Tonghe Wang, Yining Feng, Joel Beaudry, David Aramburu Nunez, Daniel Gorovets, Marisa Kollmeier, and Antonio L Damato. Instant plan quality prediction on transrectal ultrasound for high-dose-rate prostate brachytherapy. *Brachytherapy*, 24(1):171–176, 2025.

Qi Zeng, Golnoosh Samei, Davood Karimi, Claudia Kesch, Sara S Mahdavi, Purang Abolmaesumi, and Septimiu E Salcudean. Prostate segmentation in transrectal ultrasound using magnetic resonance imaging priors. *International journal of computer assisted radiology and surgery*, 13(6):749–757, 2018.

Appendix A. Prostate Segmentation Performance

This section presents the slice-level results for prostate segmentation on the in-house GOC dataset in table 8.

Table 8: Slice-level performance metrics (95% bootstrap CI) and percentage of trainable parameters, for different MedSAM fine-tuning strategies on GOC prostate segmentation. The best scores are highlighted in bold, and the second best are underlined.

Model Variant	Trainable %	Dice	Jaccard	Precision	Recall
Full Retraining	100.00	0.938 ± 0.004	0.886 ± 0.007	0.944 ± 0.011	0.936 ± 0.011
Partial IE	19.47	0.934 ± 0.005	0.880 ± 0.009	0.945 ± 0.011	0.928 ± 0.013
Freeze IE	4.34	0.923 ± 0.006	0.871 ± 0.012	0.943 ± 0.011	0.922 ± 0.014
MD Only	4.33	0.932 ± 0.006	0.875 ± 0.010	0.932 ± 0.013	$\underline{0.936 \pm 0.008}$
Selective Layers	4.11	0.932 ± 0.005	0.875 ± 0.009	0.933 ± 0.11	0.935 ± 0.009
Hybrid LoRA	0.82	$\underline{0.939 \pm 0.004}$	$\underline{0.887 \pm 0.008}$	0.947 ± 0.010	0.934 ± 0.009
MD Hypernetwork MLP	0.74	0.907 ± 0.006	0.832 ± 0.009	0.904 ± 0.009	0.914 ± 0.007
MD LoRA	0.17	0.938 ± 0.004	0.885 ± 0.009	$\mathbf{0.948 \pm 0.007}$	0.931 ± 0.011
MD Transformer LoRA	0.09	$\mathbf{0.941 \pm 0.004}$	$\mathbf{0.890 \pm 0.006}$	0.944 ± 0.006	$\mathbf{0.940 \pm 0.006}$
MD MLP LoRA	0.08	0.930 ± 0.004	0.871 ± 0.006	$\underline{0.947 \pm 0.008}$	0.917 ± 0.011
MedSAM Original	0.00	0.801 ± 0.017	0.676 ± 0.022	0.932 ± 0.006	0.714 ± 0.024
nnU-Net	-	0.836 ± 0.003	0.748 ± 0.004	0.823 ± 0.003	0.864 ± 0.003

Performance variability of all the MedSAM variants fine-tuned across the different seeds used for dataset splits is shown in figure 7.

Appendix B. Urethra Segmentation Performance

This section presents the slice-level results for urethra segmentation on the in-house GOC dataset in table 9.

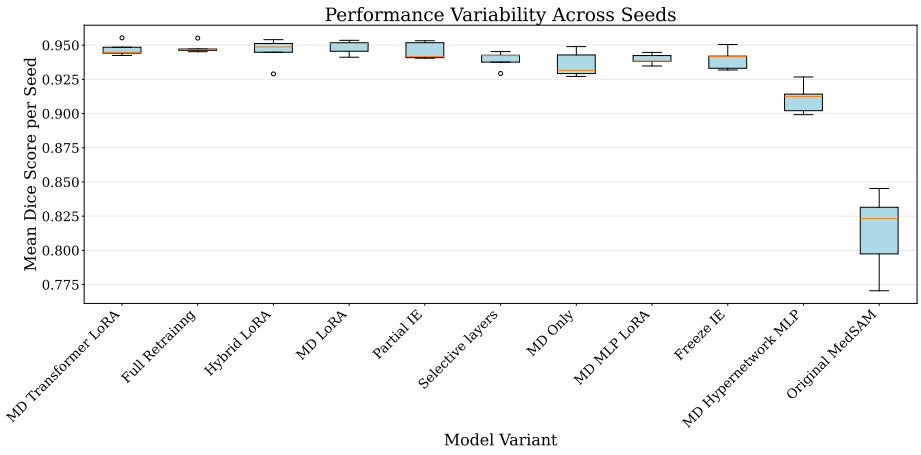

Figure 7: Performance variability of all the MedSAM variants fine-tuned across the different seeds used for dataset splits.

Table 9: Slice-level performance metrics (95% bootstrap CI) for different model variants for urethra segmentation on the in-house GOC dataset. The best scores are highlighted in bold, and the second best are underlined.

Model Variant	**Trainable %**	**Dice**	**Jaccard**	**Precision**	**Recall**
Full Retraining	100.00	0.9778 ± 0.0017	0.9570 ± 0.0030	0.9793 ± 0.0029	0.9767 ± 0.0018
Partial IE	19.47	0.9793 ± 0.0014	0.9598 ± 0.0025	0.9793 ± 0.0021	0.9795 ± 0.0012
Freeze IE	4.34	0.9780 ± 0.0015	0.9575 ± 0.0028	0.9751 ± 0.0026	$\mathbf{0.9815 \pm 0.0016}$
MD Only	4.33	0.9791 ± 0.0018	0.9594 ± 0.0032	0.9801 ± 0.0027	0.9785 ± 0.0016
Selective Layers	4.11	0.9785 ± 0.0021	0.9584 ± 0.0038	0.9772 ± 0.0040	$\underline{0.9803 \pm 0.0016}$
Hybrid LoRA	0.82	0.9796 ± 0.0013	0.9602 ± 0.0015	0.9795 ± 0.0021	0.9801 ± 0.0017
MD Hypernetwork MLP	0.74	0.9027 ± 0.0026	0.8238 ± 0.0040	0.8828 ± 0.0041	0.9272 ± 0.0042
MD LoRA	0.17	$\underline{0.9800 \pm 0.0010}$	$\underline{0.9610 \pm 0.0018}$	$\underline{0.9821 \pm 0.0017}$	0.9782 ± 0.0014
MD Transformer LoRA	0.09	$\mathbf{0.9804 \pm 0.0008}$	$\mathbf{0.9616 \pm 0.0014}$	$\mathbf{0.9828 \pm 0.0015}$	0.9781 ± 0.0016
MD MLP LoRA	0.08	0.9775 ± 0.0008	0.9563 ± 0.0015	0.9764 ± 0.0013	0.9790 ± 0.0015
Original MedSAM	0.00	0.6792 ± 0.0079	0.5226 ± 0.0095	0.9765 ± 0.0021	0.5318 ± 0.0104

Appendix C. Evaluation Results on Public Datasets

This section presents the detailed evaluation results on the three public datasets after fine-tuning on GOC dataset.

Table 10: Slice metrics for different training strategies (best in **bold**, second-best underlined). Prostate segmentation on 3 patient dataset.

Model	Trainable %	Dice	Jaccard	Dice Median	Precision	Recall
Full retraining	100.00	0.8214 ± 0.0635	0.7018 ± 0.0897	0.8269	0.9095 ± 0.0963	0.7700 ± 0.1216
Partial IE	19.47	0.8267 ± 0.0814	0.7124 ± 0.1121	0.8493	0.9368 ± 0.0511	0.7486 ± 0.1185
Freeze IE	4.34	0.7742 ± 0.1421	0.6499 ± 0.1595	0.8137	0.8873 ± 0.0739	0.8873 ± 0.0739
MD only	4.33	0.8225 ± 0.0585	0.7026 ± 0.0820	0.8390	0.8129 ± 0.1309	0.8663 ± 0.1177
Selective layers	4.11	0.8294 ± 0.0794	0.7158 ± 0.1081	0.8533	0.8081 ± 0.1251	0.8712 ± 0.0847
Hybrid LoRA	0.82	0.8488 ± 0.0448	0.7399 ± 0.0658	**0.8621**	0.8407 ± 0.0794	0.8662 ± 0.0697
MD Hypernetwork MLP	0.74	0.8473 ± 0.0472	0.7379 ± 0.0701	0.8544	0.9305 ± 0.0572	0.7867 ± 0.0897
MD LoRA	0.17	0.8443 ± 0.0554	0.7344 ± 0.0820	0.8529	0.8410 ± 0.1042	0.8653 ± 0.0913
MD Transformer LoRA	0.09	0.8551 ± 0.0473	0.7499 ± 0.0713	0.8606	0.8735 ± 0.0854	0.8498 ± 0.0870
MD MLP LoRA	0.08	$\mathbf{0.8570 \pm 0.0455}$	$\mathbf{0.7526 \pm 0.0707}$	0.8592	0.8316 ± 0.0709	$\mathbf{0.8906 \pm 0.0641}$
MedSAM Original	0.00	0.7953 ± 0.1236	0.6758 ± 0.1539	0.8277	$\mathbf{0.9774 \pm 0.0295}$	0.6878 ± 0.1610

Table 11: Volume metrics for different training strategies (best in **bold**, second-best underlined). Prostate segmentation on 3 patient dataset. (Fedorov et al., 2015)

Model	Trainable %	Dice	Jaccard	HD95	Precision	Recall
Full retraining	100.00	0.8254 ± 0.0237	0.7034 ± 0.0348	71.41 ± 63.64	0.8743 ± 0.0679	0.7906 ± 0.0661
Partial IE	19.47	0.8370 ± 0.0460	0.7224 ± 0.0663	4.72 ± 0.07	0.9350 ± 0.0095	0.7599 ± 0.0677
Freeze IE	4.34	0.7658 ± 0.1148	0.6339 ± 0.1432	20.07 ± 14.93	0.8651 ± 0.0240	0.7136 ± 0.1820
MD only	4.33	0.8126 ± 0.0510	0.6874 ± 0.0704	123.46 ± 150.32	0.7415 ± 0.0711	0.9019 ± 0.0266
Selective layers	4.11	0.8085 ± 0.0797	0.6857 ± 0.1078	106.52 ± 137.69	0.7510 ± 0.1038	0.8796 ± 0.0423
Hybrid LoRA	0.82	0.8478 ± 0.0160	0.7362 ± 0.0240	9.87 ± 3.25	0.8232 ± 0.0178	0.8742 ± 0.0187
MD Hypernetwork MLP	0.74	0.8538 ± 0.0085	0.7450 ± 0.0129	5.40 ± 1.89	0.9158 ± 0.0198	0.8001 ± 0.0127
MD LoRA	0.17	0.8444 ± 0.0138	0.7310 ± 0.0207	15.17 ± 1.29	0.8045 ± 0.0318	0.8900 ± 0.0187
MD Transformer LoRA	0.09	0.8605 ± 0.0037	0.7552 ± 0.0057	9.02 ± 2.58	0.8556 ± 0.0194	0.8665 ± 0.0218
MD MLP LoRA	0.08	$\mathbf{0.8620 \pm 0.0063}$	$\mathbf{0.7576 \pm 0.0098}$	9.09 ± 2.30	0.8254 ± 0.0097	$\mathbf{0.9023 \pm 0.0127}$
MedSAM Original	0.00	0.8362 ± 0.0349	0.7200 ± 0.0526	$\mathbf{4.08 \pm 1.54}$	$\mathbf{0.9800 \pm 0.0071}$	0.7315 ± 0.0584

Table 12: Slice metrics for different training strategies (best in **bold**, second-best underlined). Prostate segmentation on sampled dataset. (Natarajan et al., 2020)

Model	Trainable %	Dice	Jaccard	Dice Median	Precision	Recall
Full retraining	100.00	0.845 ± 0.0636	0.7363 ± 0.0878	0.8535	0.8158 ± 0.0926	0.8882 ± 0.0836
Partial IE	19.47	0.8271 ± 0.0659	0.7104 ± 0.0925	0.8310	0.8243 ± 0.0945	0.8394 ± 0.0774
Freeze IE	4.34	0.8091 ± 0.0698	0.6850 ± 0.0962	0.8136	0.7855 ± 0.0981	0.8458 ± 0.0872
MD only	4.33	0.8151 ± 0.0658	0.6930 ± 0.0923	0.8190	0.8033 ± 0.0964	0.8379 ± 0.0817
Selective layers	4.11	0.8321 ± 0.0652	0.7175 ± 0.0911	0.8383	0.8178 ± 0.0972	0.8575 ± 0.0729
Hybrid LoRA	0.82	0.8435 ± 0.0516	0.7327 ± 0.0758	0.8471	0.7923 ± 0.0843	0.9116 ± 0.0668
MD Hypernetwork MLP	0.74	0.8549 ± 0.0548	0.7504 ± 0.0809	0.8657	0.8082 ± 0.0837	$\mathbf{0.9168 \pm 0.0693}$
MD LoRA	0.17	0.8497 ± 0.0511	0.7420 ± 0.0753	0.8573	0.8028 ± 0.0814	0.9112 ± 0.0642
MD Transformer LoRA	0.09	0.8462 ± 0.0505	0.7367 ± 0.0746	0.8517	0.8018 ± 0.0833	0.9045 ± 0.0602
MD MLP LoRA	0.08	$\mathbf{0.8675 \pm 0.0479}$	$\mathbf{0.7691 \pm 0.0731}$	**0.8729**	0.8322 ± 0.0748	0.9128 ± 0.0620
MedSAM Original	0.00	0.8201 ± 0.0960	0.7052 ± 0.1250	0.8413	$\mathbf{0.9321 \pm 0.0605}$	0.7464 ± 0.1360

Table 13: Volume metrics for different training strategies (best in **bold**, second-best underlined). Prostate segmentation on sampled dataset. (Natarajan et al., 2020)

Model	Trainable %	Dice	Jaccard	HD95	Precision	Recall
Full retraining	100.00	0.8508 ± 0.0270	0.7414 ± 0.0414	64.67 ± 12.62	0.8027 ± 0.0412	0.9063 ± 0.0204
Partial IE	19.47	0.8326 ± 0.0291	0.7143 ± 0.0434	61.86 ± 9.19	0.8206 ± 0.0405	0.8461 ± 0.0291
Freeze IE	4.34	0.8067 ± 0.0323	0.6772 ± 0.0463	76.47 ± 10.71	0.7686 ± 0.0423	0.8499 ± 0.0314
MD only	4.33	0.8112 ± 0.0352	0.6839 ± 0.0509	70.35 ± 12.13	0.7872 ± 0.0455	0.8382 ± 0.0387
Selective layers	4.11	0.8358 ± 0.0316	0.7192 ± 0.0476	60.09 ± 8.20	0.8176 ± 0.0415	0.8561 ± 0.0354
Hybrid LoRA	0.82	0.8460 ± 0.0176	0.7290 ± 0.0262	71.56 ± 8.22	0.7752 ± 0.0266	0.9245 ± 0.0176
MD Hypernetwork MLP	0.74	0.8634 ± 0.0179	0.7600 ± 0.0278	66.53 ± 5.29	0.8070 ± 0.0293	$\mathbf{0.9292 \pm 0.0198}$
MD LoRA	0.17	0.8515 ± 0.0164	0.7417 ± 0.0250	67.47 ± 9.24	0.7910 ± 0.0259	0.9228 ± 0.0197
MD Transformer LoRA	0.09	0.8455 ± 0.0196	0.7328 ± 0.0292	71.04 ± 9.02	0.7888 ± 0.0323	0.9122 ± 0.0220
MD MLP LoRA	0.08	$\mathbf{0.8735 \pm 0.0202}$	$\mathbf{0.7760 \pm 0.0321}$	50.16 ± 6.84	0.8319 ± 0.0306	0.9202 ± 0.0200
MedSAM Original	0.00	0.8506 ± 0.0120	0.7402 ± 0.0180	$\mathbf{38.41 \pm 3.35}$	$\mathbf{0.9338 \pm 0.0136}$	0.7814 ± 0.0214

Table 14: Slice-level metrics for different training strategies (best in **bold**, second-best underlined). Prostate segmentation on 72 patients' dataset. (Baum et al., 2023)

Model	Trainable %	Dice	Jaccard	Precision	Recall
Full retraining	100.00	0.7777 ± 0.0802	0.6428 ± 0.1013	0.7732 ± 0.1186	0.7989 ± 0.0924
Partial IE	19.47	0.7482 ± 0.0815	0.6038 ± 0.0967	0.7904 ± 0.1101	0.7271 ± 0.1099
Freeze IE	4.34	0.7722 ± 0.0826	0.6358 ± 0.1031	0.7653 ± 0.1173	0.7987 ± 0.1053
MD only	4.33	0.7892 ± 0.0787	0.6583 ± 0.0999	0.7991 ± 0.1159	0.7964 ± 0.0948
Selective layers	4.11	0.8090 ± 0.0696	0.6846 ± 0.0918	0.8197 ± 0.1009	0.8097 ± 0.0836
Hybrid LoRA	0.82	0.8367 ± 0.0592	0.7236 ± 0.0840	0.7871 ± 0.0947	$\mathbf{0.9035 \pm 0.0601}$
MD Hypernetwork MLP	0.74	0.8468 ± 0.0562	0.7382 ± 0.0803	0.8674 ± 0.0864	0.8362 ± 0.0778
MD LoRA	0.17	0.8368 ± 0.0600	0.7237 ± 0.0846	0.8038 ± 0.1029	0.8850 ± 0.0638
MD Transformer LoRA	0.09	0.8199 ± 0.0650	0.6997 ± 0.0899	0.7907 ± 0.1065	0.8638 ± 0.0673
MD MLP LoRA	0.08	$\mathbf{0.8544 \pm 0.0482}$	$\mathbf{0.7487 \pm 0.0706}$	0.8477 ± 0.0843	0.8701 ± 0.0649
MedSAM Original	0.00	0.7737 ± 0.0944	0.6400 ± 0.1191	$\mathbf{0.9097 \pm 0.0795}$	0.6863 ± 0.1301

Table 15: Volume-level metrics for different training strategies (best in **bold**, second-best underlined). Prostate segmentation on 72 patients' dataset. (Baum et al., 2023)

Model	Trainable %	Dice	Jaccard	HD95	Precision	Recall
Full retraining	100.00	0.7818 ± 0.0496	0.6445 ± 0.0661	92.91 ± 33.85	0.7690 ± 0.0682	0.7993 ± 0.0582
Partial IE	19.47	0.7504 ± 0.0469	0.6028 ± 0.0611	101.44 ± 39.18	0.7873 ± 0.0518	0.7213 ± 0.0711
Freeze IE	4.34	0.7734 ± 0.0554	0.6338 ± 0.0730	101.96 ± 30.57	0.7562 ± 0.0719	0.7986 ± 0.0748
MD only	4.33	0.7940 ± 0.0500	0.6612 ± 0.0679	77.25 ± 25.19	0.8024 ± 0.0606	0.7897 ± 0.0660
Selective layers	4.11	0.8139 ± 0.0423	0.6883 ± 0.0589	66.47 ± 21.22	0.8230 ± 0.0547	0.8077 ± 0.0525
Hybrid LoRA	0.82	0.8427 ± 0.0287	0.7293 ± 0.0425	73.28 ± 18.92	0.7889 ± 0.0464	$\mathbf{0.9070 \pm 0.0319}$
MD Hypernetwork MLP	0.74	0.8560 ± 0.0207	0.7488 ± 0.0314	40.43 ± 17.17	0.8746 ± 0.0364	0.8397 ± 0.0303
MD LoRA	0.17	0.8443 ± 0.0256	0.7314 ± 0.0380	57.36 ± 19.64	0.8105 ± 0.0534	0.8849 ± 0.0332
MD Transformer LoRA	0.09	0.8225 ± 0.0300	0.6995 ± 0.0432	82.89 ± 22.92	0.7911 ± 0.0565	0.8606 ± 0.0388
MD MLP LoRA	0.08	$\mathbf{0.8593 \pm 0.0224}$	$\mathbf{0.7540 \pm 0.0343}$	43.09 ± 15.70	0.8532 ± 0.0428	0.8679 ± 0.0333
MedSAM Original	0.00	0.7931 ± 0.0382	0.6588 ± 0.0508	$\mathbf{26.88 \pm 7.13}$	$\mathbf{0.9031 \pm 0.0294}$	0.7087 ± 0.0518